REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* JUL 08	**2. REPORT TYPE** Conference Paper Postprint	**3. DATES COVERED** *(From - To)* 6 Jul 08 – 11 Jul 08

4. TITLE AND SUBTITLE

RANDOM CODING BOUNDS FOR DNA CODES BASED ON FIBONACCI ENSEMBLES OF DNA SEQUENCES

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)

A. D’yachkov, A. Macula, T. Renz and V. Rykov

5d. PROJECT NUMBER
232T

5e. TASK NUMBER
DN

5f. WORK UNIT NUMBER
IH

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

AFRL/RITC
525 Brooks Rd.
Rome NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR
875 N. Randolph St.
Arlington, VA 22203

AFRL/RITC
525 Brooks Rd.
Rome, NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2009-2

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited. PA# WPAFB 07-0568

13. SUPPLEMENTARY NOTES
(c) 2008 IEEE. Paper was published in the Proceedings of the 2008 IEEE International Symposium on Information Theory in Toronto, Canada, Jul 08. This work is copyrighted. One or more of the authors is a US Government employee working within the scope of their Government job; therefore, the US Government is joint owner of the work and has the right to copy, distribute, and use the work. All other rights are reserved by the copyright owner.

14. ABSTRACT
We consider the concept of a weighted 2-stem similarity function between two DNA sequences and discuss DNA codes based on the similarity. An optimal construction of such codes is suggested. A random coding bound on the rate of DNA codes is proved. To obtain the bound, we use some ensembles of DNA sequences which are generalizations of the Fibonacci sequences.

15. SUBJECT TERMS
DNA Codes, Fibonacci Ensembles, DNA Computing, Code Optimization

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** THOMAS RENZ
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	6	**19b. TELEPHONE NUMBER (*Include area code)*** N/A

Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences[0]

A. D'yachkov*, A. Macula†, T. Renz† and V. Rykov‡
*Department of Probability Theory, Faculty of Mechanics and Mathematics
Moscow State University, Moscow, 119992, Russia, Email: agd-msu@yandex.ru
†Air Force Res. Lab., IFTC, Rome Research Site, Rome NY 13441, USA,
Email: macula@geneseo.edu, thomas.renz@rl.af.mil
‡Department of Mathematics, University of Nebraska at Omaha,
6001 Dodge St., Omaha, NE 68182-0243, USA, E-mail: vrykov@mail.unomaha.edu

Abstract—**We consider the concept of a weighted 2-stem similarity function between two DNA sequences and discuss DNA codes based on this similarity. An optimal construction of such codes is suggested. A random coding bound on the rate of DNA codes is proved. To obtain the bound, we use some ensembles of DNA sequences which are generalizations of the Fibonacci sequences.**

I. Introduction

In order to accomplish DNA computing, it is necessary to have DNA libraries, also known as DNA codes, of large size and small energies of hybridization between the DNA sequences. The ultimate criterion for the value of a similarity for DNA codes is the degree to which it approximates actual bonding energies, which in turn determines the degree to which similarity approximates the likelihood of one codeword mistakenly binding to the reverse complement of another codeword. We can use a branch of mathematics known as coding theory, that was initiated around the same time that the structure of DNA was discovered, to study the space of DNA sequences endowed with a measure of similarity. The introduced measure of similarity between DNA sequences has an immediate application in determining the similarities between genes, expressed as DNA sequences, in any existing genome. Codes built on spaces of DNA sequences can be implemented in Biomolecular Computing and could have other important applications. A conventional similarity function for measuring codeword similarity is the well known deletion similarity, i.e., the length of a longest common subsequence [7]. The works of D'yachkov et al. [2], [3], [4] suggest to use the length of a longest common block subsequence, which imposes an additional adjacency requirement, with the goal of modeling actual bonding energies. In this paper, we introduce the concept of a stem similarity function which provides a more accurate estimation [1], [2] of the hybridization energy.

II. Statement of Problem

A. Notations and Auxiliary Definitions

The symbol $\triangleq$ denotes definitional equalities and the symbol $[n] \triangleq \{1,2,\ldots,n\}$ denotes the set of integers from 1 to n. Let $\{A,C,G,T\}$ be the standard DNA alphabet. For any letter $x \in \{A,C,G,T\}$, we define

$$\bar{x} \triangleq \begin{cases} T & \text{if } x = A, \\ G & \text{if } x = C, \\ C & \text{if } x = G, \\ A & \text{if } x = T \end{cases}$$

which is called a *complement* of the letter x. This means that the DNA alphabet $\{A,C,G,T\}$ consists of *two pairs of mutually complementary letters*: $\bar{A} = T$, $\bar{T} = A$ and $\bar{C} = G$, $\bar{G} = C$.

Let $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, where $x, y \in \{A,C,G,T\}^n$, be two arbitrary DNA n-sequences. By symbol $z = (z_1, z_2, \ldots, z_\ell) \in \{A,C,G,T\}^\ell$, $\ell \in [n]$, we will denote a *common subsequence* [7] of length $|z| \triangleq \ell$ between x and y. The *empty* subsequence z of length $|z| \triangleq 0$ is a common subsequence between any sequences x and y.

Definition 1. Let $2 \le r \le n$ be an arbitrary integer. A fixed DNA r-sequence $\mathbf{a} = (a_1, a_2, \ldots, a_r) \in \{A,C,G,T\}^r$, is called a *common block for sequences x and y* (briefly, *common (x,y)-block*) of length r if sequences x and y (simultaneously) contain $\mathbf{a}$ as a subsequence consisting of r consecutive elements of x and y. We will say that a common (x,y)-block $\mathbf{a}$ *yields* $r-1$ *common 2-stems* a_i, a_{i+1}, $i \in [r-1]$, containing 2 adjacent symbols of the given common (x,y)-block.

Definition 2. Let $2 \le \ell \le n$ be an integer. A sequence $z = (z_1, z_2, \ldots, z_\ell) \in \{A,C,G,T\}^\ell$ is called a *common block subsequence* of length $|z| \triangleq \ell$ between x and y if z is an *ordered collection* of non-overlapping (separated) common (x,y)-blocks and the length of each common (x,y)-block in this collection is ≥ 2. Let $\mathcal{Z}(x,y)$ be the set of all common block subsequences between x and y. For any $z \in \mathcal{Z}(x,y)$, we denote by $k(z,x,y)$, $1 \le k(z,x,y) \le |z|/2$, the *minimal number* of common (x,y)–blocks which *constitute* the given subsequence z.

Note that the difference $|z| - k(z,x,y)$, $z \in \mathcal{Z}(x,y)$, is a total number of common 2-stems containing adjacent symbols in common (x,y)-blocks constituting $z \in \mathcal{Z}(x,y)$.

Definition 3. [1], [2] For sequences $x, y \in \{A,C,G,T\}^n$, the number

$$S(x,y) \triangleq \max_{z \in \mathcal{Z}(x,y)} \{|z| - k(z,x,y)\}, \; S(x,y) \ge 0, \qquad (1)$$

[0] The work was supported by AFOSR – FA8750-07-C-0089

is called an 2-*stem similarity between* x *and* y. Obviously, $S(x,y) = S(y,x) \leq S(x,x) = n-1$.

For any $x = (x_1, x_2, \ldots, x_n) \in \{A,C,G,T\}^n$, we introduce its *reverse complement* (Watson-Crick transformation)

$$\widetilde{x} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in \{A,C,G,T\}^n. \qquad (2)$$

If $y \triangleq \widetilde{x}$, then $x = \widetilde{y}$ for any $x \in \{A,C,G,T\}^n$. If $x = \widetilde{x}$, then x is called a *self reverse complementary* sequence. If $x \neq \widetilde{x}$, then a pair $(x\,,\,\widetilde{x})$ is called a *pair of mutually reverse complementary* sequences.

Example. Let $n = 10$ and

$$x = (A, T, \underbrace{T, A, A}, A, A, \underbrace{T, T, A}),$$

$$y \triangleq \widetilde{x} = (\underbrace{T, A, A}, T, T, \underbrace{T, T, A}, A, T).$$

A common block subsequence z between x and $y = \widetilde{x}$ is

$$z \triangleq (\overbrace{T, A, A}, \overbrace{T, T, A}) = \widetilde{z} = (x_3, x_4, x_5, x_8, x_9, x_{10}) =$$

$$= (y_1, y_2, y_3, y_6, y_7, y_8) \in \mathcal{Z}(x,y).$$

The value $k(z,x,y) = 2$ and the corresponding 2–stem similarity is

$$S(x,y) \triangleq \max_{z \in \mathcal{Z}(x,y)} \{|z| - k(z,x,y)\} = 6 - 2 = 4.$$

The maximal value is achieved for the above self reverse complementary sequence $z \in \mathcal{Z}(x,y)$.

B. Weighted Stem Similarity and Distance

Let $w = w(a,b) \geq 0$, $a,b \in \{A,C,G,T\}$, be a weight function such that

$$w(a,b) = w(\bar{b}, \bar{a}), \quad a,b \in \{A,C,G,T\}. \qquad (3)$$

Condition (3) means that $w(a,b)$ is an invariant function under Watson-Crick transformation.

Definition 4. [1], [2] Let $z \in \mathcal{Z}(x,y)$ have the form

$$z \triangleq \left(z^1, z^2, \ldots, z^{k(z,x,y)}\right),$$

$$|z| = \sum_{m=1}^{k(z,x,y)} |z^m| = \sum_{m=1}^{k(z,x,y)} r_m$$

where

$$z^m \triangleq \left(z_1^m, z_2^m, \ldots, z_{r_m}^m\right) \in \{A,C,G,T\}^{r_m},$$

$$m = 1, 2, \ldots, k(z,x,y),$$

is an ordered collection of common (x,y)-blocks constituting z and $r_m \triangleq |z^m| \geq 2$ is the length of block z^m. For DNA sequences $x,y \in \{A,C,G,T\}^n$, the number

$$\mathcal{S}^{(w)}(x,y) \triangleq \max_{z \in \mathcal{Z}(x,y)} \left\{ \sum_{m=1}^{k(z,x,y)} \sum_{i=1}^{r_m - 1} w\left(z_i^m\,,\, z_{i+1}^m\right) \right\} \quad (4)$$

is called a *weighted* 2-*stem similarity between* x *and* y. We will say that $\mathcal{S}^{(w)}(x,y) \triangleq 0$ if and only if the set $\mathcal{Z}(x,y) = \varnothing$.

Function $\mathcal{S}^{(w)}(x, \widetilde{y})$ is used to model [2], [3], [4] a *thermodynamic similarity* (*hybridization energy*) between DNA sequences x and y.

Proposition 1. *For any* $x,y \in \{A,C,G,T\}^n$*, the function*

$$\mathcal{S}^{(w)}(x,y) = \mathcal{S}^{(w)}(y,x) \leq \mathcal{S}^{(w)}(x,x) \qquad (5)$$

In addition,

$$\mathcal{S}^{(w)}(x, \widetilde{y}) = \mathcal{S}^{(w)}(y, \widetilde{x}), \qquad x,y \in \{A,C,G,T\}^n. \quad (6)$$

The symmetry property and inequality (5) are evident. Equality (6) follows from definitions (2),(4) and condition (3). Identity (6) means the symmetry property of hybridization energy between DNA sequences x and y [2], [4].

One can easily check that 2-stem similarity $S(x,y)$ from Definition 3 corresponds to the uniform weight function: $w(a,b) \equiv 1$ for any $a,b \in \{A,C,G,T\}$. Table 1 shows an example [2] of values for $w(a,b)$ which satisfy (3) and have a significant biological motivation:

$w(a,b)$	$b=A$	$b=C$	$b=G$	$b=T$
$a=A$	1.02	1.46	1.29	0.88
$a=C$	1.46	1.83	2.17	1.29
$a=G$	1.32	2.24	1.83	1.46
$a=T$	0.60	1.32	1.46	1.02

Table 1.

Definition 5. [1] The number

$$\mathcal{D}^{(w)}(x,y) \triangleq \mathcal{S}^{(w)}(x,x) - \mathcal{S}^{(w)}(x,y) \qquad (7)$$

is called a *weighted* 2-*stem distance between* x *and* y.

Typically, $\mathcal{D}^{(w)}(x,y) \neq \mathcal{D}^{(w)}(y,x)$, i.e., function (7) is not symmetric. Proposition 1 gives:

$$\mathcal{D}^{(w)}(x,y) \geq \mathcal{D}^{(w)}(x,x) = 0. \qquad (8)$$

C. DNA Codes based on Stem Similarity

Let $x(j) \triangleq (x_1(j), x_2(j), \ldots, x_n(j)) \in \{A,C,G,T\}^n$, $j \in N$, be *codewords* of a *code* $X = \{x(1), x(2), \ldots, x(N)\}$ of *length* n and *size* N, where $N = 2, 4, \ldots$ is an even integer. Let D, $0 < D \leq \max\limits_{x} \mathcal{S}^{(w)}(x,x)$, be an arbitrary positive number. Taking into account (7) and (8), we give

Definition 6. A code X is called a DNA (n, D, w)-*code based on weighted* 2-*stem similarity* $\mathcal{S}^{(w)}(x,y)$ (briefly, (n, D, w)-code) if the following two conditions are fulfilled. (i). For any number $j \in [N]$ there exists $j' \in [N]$, $j' \neq k$, such that $x(j') = \widetilde{x(j)} \neq x(j)$. In other words, X is a collection of $N/2$ pairs of mutually reverse complementary sequences. (ii). For any $j, j' \in [N]$, where $j \neq j'$, the distance $\mathcal{D}^{(w)}(x(j), x(j')) \geq D$.

The following statement is obvious.

Proposition 2. *Let* (3) *be the uniform weight function, i.e.,*

$$w(a,b) \equiv 1, \qquad a,b \in \{A,C,G,T\}.$$

The corresponding symmetric distance function $\mathcal{D}^{(\equiv 1)}(x,y)$, $x,y \in \{A,C,G,T\}^n$ *has the form*

$$\mathcal{D}^{(\equiv 1)}(x,y) = \mathcal{D}^{(\equiv 1)}(y,x) = (n-1) - S(x,y), \quad (9)$$

where 2*-stem similarity* $S(x,y)$ *is defined by* (1)*, and the definition of DNA* $(n,D,\equiv 1)$*-code,* $0 < D \le n-1$*, is identified by inequality*

$$S(x(j),x(j')) \le (n-1) - D, \quad j,j' \in [N],\ j \ne j'. \tag{10}$$

Definition 7. Let $N^{(w)}(n,D)$ be the *maximal* size of DNA (n,D,w)-codes based on weighted 2-stem similarity. If $d > 0$ is a fixed number, then

$$R^{(w)}(d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_4 N^{(w)}(n,nd)}{n} \tag{11}$$

is called a *rate* of DNA (n,nd,w)-codes for a *distance fraction* d.

D. Construction of DNA $(n,2,\equiv 1)$*-codes*

In papers [3], [4], we introduced the following definitions. 1) A common subsequence $z = (z_1,\dots,z_\ell)$, $2 \le \ell \le n$, is called a *common block subsequence* of length $|z| \triangleq \ell$ between x and y if any two consecutive elements z_m, z_{m+1}, $m = 1,2,\dots,\ell-1$, which are consecutive (separated) in x are also consecutive (separated) in y and vice versa, i.e,

$$(z_m = x_{i_m}, z_{m+1} = x_{i_m+1}) \leftrightarrow (z_m = y_{j_m}, z_{m+1} = y_{j_m+1}).$$

Let $S^\beta(x,y)$, $0 \le S^\beta(x,y) \le n$, denote the length $|z|$ of *longest* sequence occurring as a common block subsequence z between sequences x and y. The number $S^\beta(x,y) = S^\beta(y,x)$ is called a *block similarity* between x and y.
2) Let D, $1 \le D \le n-1$, be an arbitrary integer. A code X is called a DNA (n,D)-code based on block similarity $S^\beta(x,y)$ if the following two conditions are fulfilled. (i) For any number $j \in [N]$ there exists $j' \in [N]$, $j' \ne j$, such that $x(j') = \widetilde{x(j)} \ne x(j)$. (ii) For any $j,j' \in [N]$, where $j \ne j'$, the block similarity $S^\beta(x(j),x(j')) \le n-D-1$. For given n and D, we denote by $N^\beta(n,D)$ the *maximal size* of (n,D)-codes based on block similarity.

Let $x,y \in \{A,C,G,T\}^n$ be arbitrary DNA sequences. One can easily see that block similarity $S^\beta(x,y) = n-2$ iff the corresponding 2-stem similarity $S(x,y) = n-3$. Therefore, from (9)-(10) it follows that the definition of DNA $(n,1)$-code based on block similarity is equivalent to the definition of DNA $(n,2,\equiv 1)$-codes based on 2-stem similarity. This means that $N^\beta(n,1) = N^{(\equiv 1)}(n,2)$. Hence, the main result of paper [4] about constructions of optimal DNA codes based on block similarity leads to

Theorem 1. *If* $n = 4m$, $m = 1,3,5,\dots$, *then*

$$N^{(\equiv 1)}(n,2) = \frac{4^{n-1}+4}{2}.$$

E. DNA Codes for Fibonacci Ensembles

Let L be a collection of 2-strings of DNA letters, *closed under reverse complement transformation.* For instance,

$$L = \varnothing, \quad L = \{TA\}, \quad L = \{TA, AT\}$$
$$L = \{TA, AT, AA, TT\}. \tag{12}$$

Denote by $DNA(n,L)$ (briefly, $[n,L]$) the set (ensemble) of all DNA sequences which *do not contain* 2*-stems from* L. We will say that $[n,L]$ is the *Fibonacci* L*-ensemble*[1]. Denote by $\lambda_L(n) \triangleq |DNA(n,L)| = |[n,L]|$ the *cardinality* of $[n,L]$.

Definition 8. Let $N_L(n,D)$ be the *maximal size* of DNA $(n,D,\equiv 1)$-codes $X \subseteq DNA(n,L)$. If the distance fraction $d > 0$ is a fixed number, then

$$R_L(d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_4 N_L(n,nd)}{n} \tag{13}$$

is called a *rate* of DNA codes for the Fibonacci L-ensemble.

For a weight function (3), introduce numbers

$$\underline{w}_L \triangleq \min_{(a,b)\notin L} w(a,b). \tag{14}$$

For instance, if the values of $w = w(a,b)$ are given by Table 1, then

$$\underline{w}_L = \begin{cases} 0.60 & \text{if } L = \varnothing, \\ 0.88 & \text{if } L = \{TA\}, \\ 1.02 & \text{if } L = \{TA, AT\}, \\ 1.29 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases} \tag{15}$$

One can easily check [1] that the distance

$$\mathcal{D}^{(w)}(x,y) \ge \underline{w}_L \cdot \mathcal{D}^{(\equiv 1)}(x,y) \quad \text{if} \quad x,y \in DNA(n,L).$$

In virtue of (9) and (10), this gives

Proposition 3. *Let* $\underline{w}_L$ *be a number defined by* (14) *and a code* $X \subset DNA(n,L)$*. If* X *is a DNA* $(n,D,\equiv 1)$*-code, then* X *is a DNA* $(n,\underline{w}_L \cdot D, w)$*-code. Hence, rate* (11) *satisfies inequality*

$$R^{(w)}(d) \ge \max_L R_L\left(\frac{d}{\underline{w}_L}\right), \tag{16}$$

where $R_L(d)$ *is defined by* (13).

In the rest part of paper, we obtain a random coding bound on $R_L(d)$ for L defined by (12). Then applying (16), we get a random coding bound on the rate $R^{(w)}(d)$ of DNA (n,nd,w)-codes based on weighted 2-stem similarity.

III. Random Coding Bounds

A. On Cardinalities of Fibonacci L*-Ensembles*

If $L = \varnothing$, then $\lambda_L(n) = 4^n$. If $L \ne \varnothing$, then cardinalities $\lambda_L(1) = 4$ and $\lambda_L(2) = 16 - |L|$ are given. For sets L define by (12), we calculate cardinalities $\lambda_L(n)$, $n = 3,4\dots$, using the following well known result from the theory of recurrent sequences.

Proposition 4. *Let* $f_1 \ne 0$ *and* $f_2 \ne 0$ *be arbitrary fixed numbers. If sequence* $\lambda_L(n)$*,* $n = 3,4,\dots$*, satisfies recurrent equation*

$$\lambda_L(n) = f_1\,\lambda_L(n-1) + f_2\,\lambda_L(n-2), \tag{17}$$

then

$$\lambda_L(n) = C_1\,r_1^n + C_2\,r_2^n, \quad n = 1,2,\dots, \tag{18}$$

where $r_1 = r_1(L)$ *and* $r_2 = r_2(L)$ *are roots of the characteristic equation* $r^2 - f_1\,r - f_2 = 0$ *and* $C_1 = C_1(L)$*,* $C_2 = C_2(L)$

[1] Binary $0,1$-sequences which do not contain 2-stems of the form $(1,1)$ are known as the Fibonacci sequences [6].

are calculated from initial conditions: $4 = C_1 r_1 + C_2 r_2$, $16 - |L| = C_1 r_1^2 + C_2 r_2^2$.

Formula (18), obviously, leads to

Proposition 5. *If r_1, r_2 are real numbers, $r_1 > 0$ and $r_1 > |r_2|$, then $\lambda_L(n)$, $n = 1, 2, \ldots$, satisfies inequalities*

$$C\, r^n\, [1 - \beta\, \alpha^n] \le \lambda_L(n) \le C\, r^n\, [1 + \beta\, \alpha^n], \tag{19}$$

where

$$r = r_1 \triangleq \max\{r_1, r_2\}, \quad C \triangleq C_1,$$

$$\alpha \triangleq \left|\frac{r_2}{r_1}\right| < 1, \quad \beta \triangleq \left|\frac{C_2}{C_1}\right|. \tag{20}$$

Remark. For the case $L = \varnothing$, bounds (19) will be true as well (with the sign of equality) if we formally define $r_1 = 4$, $C_1 = 1$ and $r_2 = C_2 = 0$, i.e., $C = 1$, $r = 4$ and $\alpha = \beta = 0$.

Lemma 1. *If $L = \{TA\}$, then $\lambda_L(n)$ satisfies* (17), *where $f_1 = 4$, $f_2 = -1$. Hence, parameters* (20) *of bounds* (19) *are:*

$$r = 2 + \sqrt{3} = 3.73, \quad C = \frac{3 + 2\sqrt{3}}{6} = 1.08,$$

$$\alpha = \beta = 7 - 4\sqrt{3} = .0718. \tag{21}$$

Lemma 2. *If $L = \{TA, AT\}$, then $\lambda_L(n)$ satisfies* (17), *where $f_1 = 3$, $f_2 = 2$. Hence, parameters* (20) *of bounds* (19) *are:*

$$r = \frac{3 + \sqrt{17}}{2} = 3.56, \quad C = \frac{17 + 5\sqrt{17}}{34} = 1.11,$$

$$\alpha = \frac{13 - 3\sqrt{17}}{4} = .158, \quad \beta = \frac{21 - 5\sqrt{17}}{4} = .0961. \tag{22}$$

Lemma 3. *If $L = \{TA, AT, AA, TT\}$, then $\lambda_L(n)$ satisfies* (17), *where $f_1 = 2$, $f_2 = 4$. Hence parameters* (20) *of bounds* (19) *are:*

$$r = 1 + \sqrt{5} = 3.24, \quad C = \frac{5 + 3\sqrt{5}}{10} = 1.17,$$

$$\alpha = \frac{3 - \sqrt{5}}{2} = .382, \quad \beta = \frac{7 - 3\sqrt{5}}{2} = .146. \tag{23}$$

Proof of Lemmas 1-3. Let $a, b \in \{A, C, G, T\}$ denote arbitrary letters of DNA alphabet and

$$[n, L]_a \triangleq \{\, x \,:\, x \in [n, L] \quad \text{and} \quad x_n = a \}\,,$$

$$[n, L]_{a,b} \triangleq \{\, x \,:\, x \in [n, L] \quad \text{and} \quad x_{n-1} = a, x_n = b \}\,,$$

denote the corresponding subsets of ensemble $[n, L]$. If a pair $(a, b) \in L$, then subset $[n, L]_{a,b} = \varnothing$. Note that $[n, L]_a$ and $[n, L]$ can be written as sums of non-intersecting subsets:

$$[n, L]_a = [n, L]_{A,a} + [n, L]_{C,a} + [n, L]_{G,a} + [n, L]_{T,a}$$

$$[n, L] = [n, L]_A + [n, L]_C + [n, L]_G + [n, L]_T. \tag{24}$$

In addition, one can easily see the following two properties.
1) If for any $b \in \{A, C, G, T\}$, pair $(b, a) \notin L$, then the cardinality

$$|[n, L]_a| = |[n - 1, L]| = \lambda_L(n - 1). \tag{25}$$

2) For any pair $(a, b) \notin L$, the cardinality

$$|[n, L]_{a,b}| = |[n - 1, L]_a|\,. \tag{26}$$

Let $L = \{TA\}$. In virtue of (24)-(26), we have

$$\lambda_L(n) = 3\,\lambda_L(n-1) + |[n, L]_{A,A}| + |[n, L]_{C,A}| + |[n, L]_{G,A}| =$$

$$= 3\,\lambda_L(n-1) + |[n-1, L]_A| + |[n-1, L]_C| + |[n-1, L]_G| =$$

$$= 3\,\lambda_L(n-1) + 2\,\lambda_L(n-2) + |[n-1, L]_A|\,.$$

and

$$\lambda_L(n-1) = 3\,\lambda_L(n-2) + |[n-1, L]_A|\,.$$

These formulas yield the recurrent equation

$$\lambda_L(n) = 4\lambda_L(n-1) - \lambda_L(n-2), \quad n = 3, 4 \ldots,$$

formulated in Lemma 1. Using the similar arguments, one can prove Lemma 2 for set $L = \{TA, AT\}$ and Lemma 3 for set $L = \{TA, AT, AA, TT\}$.

B. Random Coding Bound for Fibonacci L-Ensemble

Let

$$\rho_L \triangleq \log_4 r, \quad \rho_L' \triangleq \log_4 \frac{r}{C^3(1 + \beta\alpha^2)(1 + \beta\alpha)^2}.$$

where $r = r(L)$, $C = C(L)$, $\alpha = \alpha(L)$ and $\beta = \beta(L)$ are introduced in Propositions 4 and 5 and given by formulas (20). For sets L defined by (12), parameters (20) are calculated by formulas (21)-(23). In Sect. IV, using a random coding method [4], we present a brief proof of

Theorem 2. *For any distance fraction $d > 0$, the rate* (13) *satisfies inequality*

$$R_L(d) \ge \underline{R}_L(d) \triangleq \min_{0 \le u \le d} \{(1 - u)\rho_L - E_L(u)\},$$

where

$$E_L(u) \triangleq \max_{0 \le v \le \min\{u,\, 1-u\}} E^L(v, u),$$

$$E^L(v, u) \triangleq -\rho_L' \cdot v + (1 - u)\, h_4 \left(\frac{v}{1 - u}\right) + 2\, u\, h_4 \left(\frac{v}{u}\right),$$

$$h(u) \triangleq -u \log_4 u - (1 - u) \log_4 (1 - u).$$

Let a number d_L, $0 < d_L < 1$, be the unique root of equation $\underline{R}_L(d) = 0$ or $(1 - d)\rho_L = E_L(d)$. Obviously, if $0 < d < d_L$, then $R_L(d) > 0$ and the following lower bound

$$R_L(d) \ge \underline{R}_L(d) \triangleq (1 - d)\rho_L - E_L(d), \qquad 0 < d < d_L,$$

holds. Function $\underline{R}_L(d)$ is called a *random coding bound* on the rate $R^L(d)$. We will say that the number d_L, $0 < d_L < 1$, is a *critical distance fraction* of the random coding bound $\underline{R}_L(d)$ for $DNA(n, L)$-ensemble.

For sets (12), our calculations based on Lemmas 1-3 give the following numerical values for critical distance fractions:

$$d_L = \begin{cases} 0.4794 & \text{if } L = \varnothing, \\ 0.4316 & \text{if } L = \{TA\}, \\ 0.4054 & \text{if } L = \{TA, AT\}, \\ 0.3487 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases} \tag{27}$$

C. Random Coding Bound for DNA (n, dn, w)-Codes

Let $R^{(w)}(d)$, $d > 0$, be the rate (11) of DNA (n, dn, w)-codes and d_L, $0 < d_L < 1$, is the critical distance fraction of random coding bound $\underline{R}_L(d)$ for Fibonacci L-ensemble. Propositions 3 and Theorem 2 lead to

Theorem 3. *If* $0 < d < d^{(w)} \triangleq \max\limits_{L}\{\underline{w}_L \cdot d_L\}$, *then the rate* $R^{(w)}(d) > 0$ *and lower bound*

$$R^{(w)}(d) \;\geq\; \underline{R}^{(w)}(d) \triangleq \max_{L} \left\{ \underline{R}_L \left(\frac{d}{\underline{w}_L} \right) \right\}$$

holds.

Function $\underline{R}^{(w)}(d)$ is called a random coding bound for DNA (n, dn, w)-codes. The number $d^{(w)} > 0$ is called a *critical distance fraction* of the random coding bound $\underline{R}^{(w)}(d)$. For instance, if weight function $w = w(a, b)$ is defined by Table 1, then for sets (12), numbers (15) and (27) give:

$$\underline{w}_L \cdot d_L = \begin{cases} 0.29 & \text{if } L = \varnothing, \\ 0.38 & \text{if } L = \{TA\}, \\ 0.41 & \text{if } L = \{TA, AT\}, \\ 0.45 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases}$$

Therefore, the corresponding critical distance fraction is $d^{(w)} \triangleq \max\limits_{L}\{\underline{w}_L \cdot d_L\} = 0.45$.

IV. Proof of Theorem 2

Let $S(x,y)$ be 2-stem similarity (1) for the uniform weight function. For an arbitrary integer $s \in [n-1]$, define the set $\mathcal{P}_L(n,s) \triangleq \{(x,y) \in [n,L] \times [n,L] \,:\, S(x,y) = s\}$.

Lemma 4. *The size*

$$|\mathcal{P}_L(n,s)| \;\leq\; \sum_{j=1}^{\min\{s,n-s\}} r^{s+j} \binom{s-1}{j-1} \left[C\,(1 \,+\, \beta\, \alpha^2) \right]^j \;\times$$

$$\times \left\{ r^{n-s-j} \left[C\,(1 \,+\, \beta\, \alpha) \right]^{j+1} \binom{n-s}{j} \right\}^2, \qquad (28)$$

where $r = r(L)$, $C = C(L)$, $\alpha = \alpha(L)$ and $\beta = \beta(L)$ were introduced in the formulation of Theorem 2.

The random coding method of [4], Lemma 4 and an asymptotic analysis on the right-hand side of (28) yield Theorem 2. To complete the proof of Theorem 2, we give

Proof of Lemma 4. Consider a pair $(x,y) \in A^n \times A^n$ for which $S(x,y) = s$. Then there exists $z \in \mathcal{Z}(x,y)$, $|z| \leq n$, and the integer $j = k(z,x,y) \leq |z|/2$ for which equalities

$$s = |z| - j \iff |z| = s + j \iff n - |z| = n - s - j$$

take place. It follows that for any $z \in \mathcal{Z}(x,y)$, the number $j = k(z,x,y)$ satisfies inequalities $1 \leq j \leq \min\{s\,;\, n-s\}$.

Obviously, the number of all ways to distribute $|z|$ indistinguishable marbles in j boxes provided that each of j boxes contains ≥ 2 marbles is $\binom{s-1}{j-1}$. In addition, the number of all ways to distribute $n - |z|$ indistinguishable marbles in $j+1$ boxes if empty boxes are accepted is $\binom{n-s}{j}$.

Let $1 \leq j \leq b \leq n$ be fixed integers and

$$\{b_\ell\} \triangleq (b_1, b_2, \ldots, b_\ell, \ldots, b_j), \quad b_\ell \geq 1,$$

is an ordered collection of integers. For $m = 1, 2$, introduce two sets

$$(\{b_\ell\})_m \triangleq \left\{ \{b_\ell\} \,:\, \sum_{\ell=1}^{j} b_\ell \,=\, b, \quad b_\ell \geq m \right\} \qquad (29)$$

and define numbers

$$\widetilde{\lambda}_L^m(j,b) \;\triangleq\; \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \lambda_L(b_\ell) \right\}. \qquad (30)$$

Applying above formulas and notations, one can see that for any $s \in [n-1]$, the cardinality

$$|\mathcal{P}_L(n,s)| \;\leq\; \sum_{j=1}^{\min\{s\,;\,n-s\}} \widetilde{\lambda}_L^2\,(j\,,\, s+j) \,\cdot\, \binom{s-1}{j-1} \times$$

$$\times \left[\widetilde{\lambda}_L^1\,(j+1\,,\, n-s-j) \binom{n-s}{j} \right]^2. \qquad (31)$$

From definition (29)-(30) and upper bound (19) it follows that for $m = 1, 2$,

$$\widetilde{\lambda}_L^m(j,b) \;\leq\; \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \left[C\, r^{b_\ell}\, \left(1 \,+\, \beta\, \alpha^{b_\ell} \right) \right] \right\} \;\leq$$

$$\leq\; C^j\, r^b \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \left[1 + \beta \alpha^{b_\ell} \right] \right\} \leq r^b \left[C(1 + \beta \alpha^m) \right]^j.$$

These inequalities and (31) lead to (28).

Lemma 4 is proved.

References

[1] *Bishop M.A.,D'yachkov A.G., Macula A.J., Renz T.E., Rykov V.V.,* Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes // Journal of Computational Biology, 2007, V. 14, N. 8, P. 1088-1104.

[2] *D'yachkov A.G.,Macula A.J.,Pogozelski W.K., Renz T.E., Rykov V.V., Torney D.C.,* A Weighted Insertion—Deletion Stacked Pair Thermodynamic Metric for DNA Codes // Proc. of 10th Int. Workshop on DNA Computing. Milan, Italy, 2004, P. 90–103.

[3] *D'yachkov A.G.,Macula A.J., Renz T.E.,Vilenkin P.A,Ismagilov I.K,* New Results on DNA Codes // Proc. of the 2005 IEEE International Symposium on Information Theory, Adelaide, South Australia, Australia, September 4 - 9, 2005, P. 283-288.

[4] *D'yachkov A.G., Macula A.J., Torney D.C., Vilenkin P.A., White P.S., Ismagilov I.K., Sarbayev R.S.,* On DNA Codes // Probl. Peredachi Informatsii, 2005, V. 41, N. 4, P. 57-77, (in Russian). English translation: Problems of Information Transmission, V. 41, N. 4, 2005, P. 349-367.

[5] *D'yachkov A.G., Erdos P.L., Macula A.J., Rykov V.V., Torney D.C., Tung C.S., Vilenkin P.A., White P.S.,* Exordium for DNA Codes // J. Comb. Optimization, V. 7, N. 4, 2003, P. 369–379.

[6] Cameron P.J., *Combinatorics: Topics, Techniques, Algorithms*, Cambridge University Press, 1994.

[7] *Levenshtein V.I.,* Efficient Reconstruction of Sequences from Their Subsequences and Supersequences // J. Comb. Th., Ser. A, V. 93, 2001, P. 310-332.